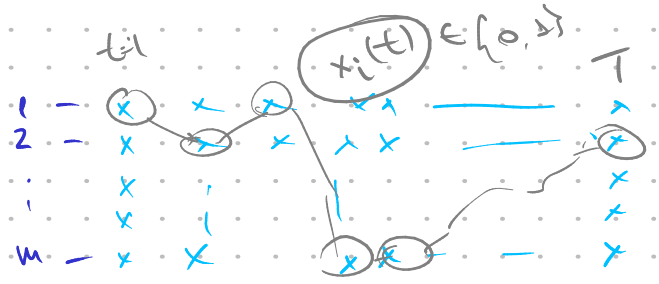
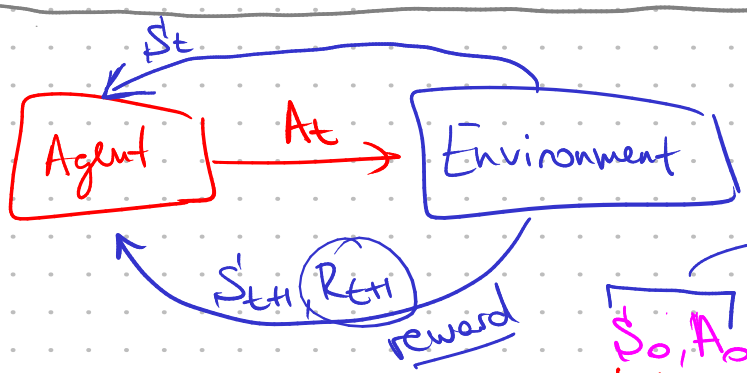


a_1 c_1 $r_1(a_1, c_1)$
a_2 c_2 $r_2(a_2, c_2)$
a_t c_t $r_t(a_t, c_t)$

A/B testing ⟶ r_0
⟶ r_1
⟶ web (m+1)-я ручка



MDP - Markov Decision Process

$p(r, s' | s, a)$

$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \dots$

π π $\pi(A_t|S_t)$

$p(r, s'|s, a)$, $p(s'|s, a)$,

$r(s,a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$

$r(s,a,s') = \mathbb{E}[R_{t+1} | S_t = s, A_t = a, S_{t+1} = s']$

① Что оптимизируем? G_t - expected return

$\gamma = 1$, $R_{T+k} = 0$

Episodic tasks: $G_t = R_{t+1} + R_{t+2} + \dots + R_T$

Continuous tasks: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

$\leq \max R \cdot \frac{1}{1-\gamma}$

$G_t = R_{t+1} + \gamma \cdot G_{t+1}$

② Value functions

- State value function

$$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_{\substack{\pi \\ A_t \sim \pi(S_t)}}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

- action value function

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

$$\pi^*(s) = \arg\max_{\pi} V_\pi(s) = \arg\max_{\pi,a} Q_\pi(s,a)$$

$$Q_\pi(s,a) = \mathbb{E}_{R_{t+1},S_{t+1},A_{t+1},R_{t+2},\dots}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s, A_t = a\right] =$$

$$= \mathbb{E}_{R_{t+1},S_{t+1}}\left[\mathbb{E}_\pi\left[R_{t+1} + \dots \mid \dots\right]\right] =$$

$$= \sum_{r,s'} p(r,s' \mid s,a)\left(r + \mathbb{E}_\pi\left[\gamma R_{t+2} + \gamma^2 R_{t+3} \dots \mid S_t = s, A_t = a, S_{t+1} = s'\right]\right)$$

$$\boxed{Q_\pi(s,a) = \sum_{r,s'} p(r,s' \mid s,a)\left(r + \gamma V_\pi(s')\right)}$$

$$V_\pi(s) = \mathbb{E}_{A_t,R_{t+1},S_{t+1},\dots}\left[G_t \mid S_t = s\right] = \sum_a \pi(a \mid s)\,\mathbb{E}_\pi[\dots]$$

$$\boxed{V_\pi(s) = \sum_a \pi(a \mid s) \cdot Q_\pi(s,a)}$$

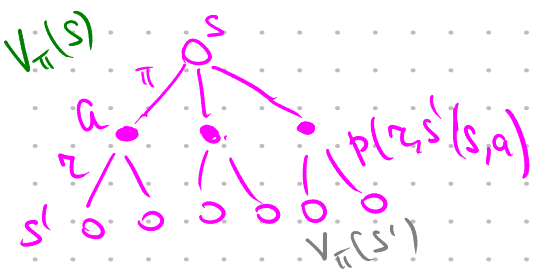
Bellman equation

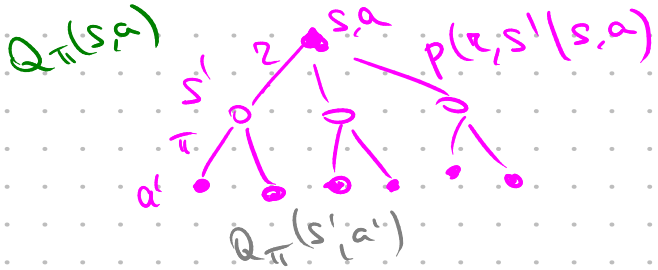
$$\boxed{V_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,s'} p(r,s' \mid s,a)\left(r + \gamma V_\pi(s')\right)}$$

$$\begin{pmatrix} V_\pi(s_1) \\ \vdots \\ V_\pi(s_N) \end{pmatrix} = \begin{pmatrix} \\ \\ \\ \end{pmatrix}\begin{pmatrix} V_\pi(s_1) \\ \vdots \\ V_\pi(s_N) \end{pmatrix} \qquad \vec{x} = f(\vec{x})$$

$$\boxed{Q_\pi(s,a) = \sum_{r,s'} p(r,s' \mid s,a)\left(r + \gamma \sum_{a'} \pi(a' \mid s')\, Q_\pi(s',a')\right)}$$

backup diagrams





$$V_*(s) = \max_\pi V_\pi(s) = \max_\pi \left(\sum_a \pi(a|s) \sum_{r,s'} p(r,s'|s,a)\left(r+\gamma V_\pi(s')\right)\right)$$

$$= \max_{\pi(a|s)} \max_\pi \left(\sum_a \pi(a|s) \sum_{r,s'} p(r,s'|s,a)\left(r+\gamma V_\pi(s')\right)\right)$$

$$= \max_a \sum_{r,s'} p(r,s'|s,a)\left(r+\gamma \max_\pi V_\pi(s')\right)$$

$$V_*(s) = \max_a \left[\sum_{r,s'} p(r,s'|s,a)\left(r+\gamma V_*(s')\right)\right]$$

$\bar{x} = f(\bar{x})$

Bellman equation

$$Q_*(s,a) = \max_\pi Q_\pi(s,a) = \max_\pi \sum_{r,s'} p(r,s'|s,a)\left(r+\gamma \cdot \sum_{a'} \pi(a'|s') Q(s',a')\right)$$

$$= \sum_{r,s'} p(r,s'|s,a)\left(r+\gamma \cdot \max_{\pi(a'|s')} \max_\pi \left[\sum_{a'} \pi(a'|s') Q(s',a')\right]\right)$$

$$Q_*(s,a) = \sum_{r,s'} p(r,s'|s,a)\left(r+\gamma \max_{a'} Q_*(s',a')\right)$$

Bellman equation

Policy improvement theorem

$$\pi_0 \to \pi_1 \to \pi_2 \to \cdots \to \pi_*$$

$\pi' \geq \pi$: $\forall s \quad V_{\pi'}(s) \geq V_\pi(s)$

Thm. Если для π и π' верно, что

$$\forall s \quad Q_\pi(s, \pi'(s)) \geq V_\pi(s),$$

то

$$\forall s \quad V_{\pi'}(s) \geq V_\pi(s)$$

Док-во:

$$V_\pi(s) \leq Q_\pi(s, \pi'(s)) = \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s\right] \leq$$

$$\leq \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma Q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\right] =$$

$$= \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 V_\pi(S_{t+2}) \mid S_t = s\right] \leq$$

$$\leq \mathbb{E}_{\pi'}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \mid S_t = s\right] = V_{\pi'}(s)$$

Policy iteration

$$\pi_{k+1}(s) = \arg\max_a Q_{\pi_k}(s, a)$$